Topoi
https://doi.org/10.1007/s11245-025-10304-2

On the Philosophical Naivety of Engineers in the Age of Machine Learning

M. Z. Naser[1,2]

Received: 4 June 2025 / Accepted: 8 October 2025
© The Author(s) 2025

Abstract
This paper examines the paradoxical decline in engagement with philosophy of science among engineers precisely when machine learning (ML) systems are increasingly performing complex epistemological functions in engineering practice. We identify how *philosophical naivety*, characterized by the uncritical adoption of reductive frameworks regarding consciousness, intelligence, and ethics, creates tangible organizational and technical liabilities. We then demonstrate how conceptual limitations in engineers' philosophical foundations lead to three primary flaws: 1) ontological misclassification of system capabilities, 2) ethical blind spots in ML system design and application, and 3) inadequate epistemological approaches and hidden philosophical commitments for interpreting model outputs. Thus, we argue that renewed engagement with the philosophy of science is not merely academic but necessary for engineers to maintain epistemic authority and responsibility in an era where engineering judgment is increasingly delegated to or mediated by ML systems. In response, we propose a technical-philosophical framework integrating perspectives from philosophy of mind, ethics, epistemology, and engineering to address these shortcomings systematically.

Keywords Engineering · Artificial intelligence · Epistemology

1 Introduction

The rise and interest of adopting machine learning (ML) systems in engineering has created a unique epistemological challenge, wherein engineers increasingly delegate critical decisions to systems whose operational logic fundamentally differs from traditional methods (Wheeler 2017). While classical engineering relies on mechanistic models grounded in physical laws and causal relationships, contemporary ML systems operate through statistical pattern recognition to produce outputs that appear rigorous (i.e., yield high predictivity) yet lack the explanatory transparency of their predecessors (Rudin 2019).

This shift has occurred precisely when engineering requires heightened philosophical sophistication to navigate questions about what constitutes knowledge, explanation, and inference in ML. To add to that, the engineering profession has moved in the opposite direction, with formal engagement with the philosophy of science diminishing in curricula and professional development programs (Mitcham 1998). This divergence between increasing epistemological complexity and decreasing philosophical literacy creates a dangerous gap in systematic misunderstandings about ML capabilities, uncritical acceptance of model outputs, and the erosion of engineering's epistemic authority (Christie and Graaff 2017; Michelfelder and Doorn 2020).

These systematic failures demand a closer examination of their root causes and broader implications. Analysis of these systematic failures reveals that the core problem lies not in numerical or implementation instability but in a deeper epistemic lapse, wherein engineers often naively treat the output of a ML model as if it possessed the causal insight of a mechanics-based simulator (Moraffah et al. 2020; Naser 2024). As Lipton (Lipton 2018a) notes, engineers often conflate different notions of explanation, understanding, and interpretation without recognizing these as distinct epistemological categories. This misclassification is grounded in tacit assumptions about how knowledge resides in data

✉ M. Z. Naser
mznaser@clemson.edu

[1] School of Civil and Environmental Engineering & Earth Sciences, Clemson University, Clemson, USA

[2] Artificial Intelligence Research Institute for Science and Engineering (AIRISE), Clemson University, Clemson, USA

Published online: 25 November 2025

and algorithms. In fact, recent empirical studies have documented significant challenges in how engineers conceptualize and reason about ML systems. In one study, Ishikawa and Yoshioka (Ishikawa and Yoshioka 2019) surveyed 278 ML practitioners and identified key perceived difficulties, including low accuracy, absence of an oracle, and uncertainty regarding system behavior, all of which suggest fundamental epistemological challenges. More recently, Kim and McGill (Kim and Im 2023) conducted an empirical study that revealed how *anthropomorphism* is a multidimensional construct involving both morphological factors and human-like intelligence, with engineers consistently attributing cognitive and emotional intelligence to ML systems despite their fundamentally different operational principles.

The tendency to apply human-centered ontological categories to ML systems has also been empirically documented across multiple domains. For example, research on shared mental models in human-ML teams shows that an anthropomorphized agent visualized as an avatar led human users to have human-like expectations from it—for better or worse, indicating systematic biases in how engineers attribute agency and intentionality to ML systems (Andrews et al. 2023). The same anthropomorphic bias in shared mental model formation occurs regardless of specific engineering background and implies a systematic tendency to apply human-centered ontological categories to ML systems. This pattern extends beyond software domains, as demonstrated in civil engineering where professionals consistently apply human-centered decision making frameworks when evaluating ML recommendations, often leading to systematic mismatches between their expectations and actual system behavior (Zhang et al. 2023).

Furthermore, research documents how engineers consistently hold *folk theories* and *algorithmic imaginaries* about ML functionality by invoking both positive and negative machine stereotypes that reflect systematic anthropomorphic reasoning rather than technical understanding. McDermott (McDermott 1976) famously complained about ML researchers' and programmers' use of *wishful mnemonics* (e.g., terms like *understand* or *learn* referred to ML), which he considered to be misleading both for researchers and for the public (Salles et al. 2020). The above empirical evidence collectively demonstrates that engineering practice inadvertently imports philosophical assumptions from human cognition (i.e., particularly ontological categories related to understanding, reasoning, mind, and agency) when theorizing about ML behavior, despite these concepts potentially misrepresenting the actual computational mechanisms underlying ML systems.

From the lens of this work, *philosophical naivety* refers to an unexamined default stance that treats intelligence as merely the successful manipulation of statistical correlations, consciousness as irrelevant to engineered cognition, ethics as an afterthought reducible to a checklist, and knowledge as whatever maximizes predictive accuracy. Such a stance is not malicious; rather, it is a pragmatic shortcut that allows engineers of various backgrounds,[1] often with little training/exposure to ML, to integrate potent algorithms without wading into conceptual debates. This shortcut becomes costly when it omits ontological differences between numerical models and physical systems or promotes epistemic complacency about how uncertainty is quantified (Cohen et al. 2021). Simply put, if engineers cannot defend the epistemic legitimacy of decisions mediated by opaque ML models, courts, insurers, and the public may reclassify certain tasks as inherently unsafe (Kotsoglou 2013). This is because licensed engineers (i.e., Professional Engineers or PEs) sign/seal designs, carry the duty of care, and are the parties deposed after failures. Courts and insurers ask whether the engineer exercised defensible professional judgment (i.e., they do not accept 'the model said so' as sufficient). Therefore, engineers must be able to justify the epistemic basis of ML-mediated decisions at least to the standard of care (traceability, reason-giving, known limits) (NSPE Code of Ethics for engineers 1993; Hoke 2020; Furey et al. 2021).

Three classes of cost follow. First, ontological misclassification arises when learning systems are implicitly granted capabilities they do not possess, such as counterfactual reasoning or domain extrapolation (Kasirzadeh and Smart 2021). The practical consequences of such misclassification include that systems designed under these assumptions exhibit brittleness when deployed in contexts that violate their implicit ontological commitments (Marcus et al. 2019). Second, ethical blind spots proliferate because optimization targets are mistaken for value-neutral metrics (which can obscure distributive consequences or safety externalities until deployment) (Barocas et al. 2023). Third, inadequate epistemological frameworks lead to brittle validation regimes that accept spurious correlations as truth (which creates a hidden technical debt that surfaces only after models encounter novel regimes) (Ye et al. 2024).

This paper proceeds as follows: Sect. 2 provides a concise historical and conceptual map of engineering's philosophical engagement, clarifies key terms, and identifies the specific gaps that motivate our study. Then, Sect. 3 catalogues three concrete failure modes of philosophical naivety, namely ontological misclassification, ethical blind spots, and inadequate epistemic frameworks. Section 4

[1] We do not confine this work to computer, software, or ML engineers. Rather, we believe our findings apply broadly to many engineering disciplines—especially those that have yet to widely adopt ML techniques at scale (e.g., civil, mechanical, etc.) or fully integrate ML into their curricula.

elevates these flaws to the organizational level and analyzes their implications for risk management, accountability, professional liability, and the erosion of engineering authority. Section 5 introduces a six-stage technical-philosophical framework that operationalizes ontology, epistemic warrant, values, interpretability, responsibility, and continuous auditing as deploy-time controls. Section 6 sets out a coordinated research agenda for industry, academia, and professional societies, including toolchains, curricula, and longitudinal studies to institutionalize the framework. Finally, Sect. 7 concludes this work.

2 The Current State of Philosophical Engagement in Engineering

The relationship between engineering and philosophy has undergone a significant transformation since the early twentieth century. Mitcham (Mitcham 1998) traces this evolution and notes that engineering emerged from craft traditions that emphasized practical knowledge over theoretical contemplation. This historical divide was reinforced by the professionalization of engineering education (which prioritized mathematical formalism and empirical validation while marginalizing philosophical inquiry) (Goldberg and Somerville 2015).

From the end of World War II through the late 1980 s, engineering functioned as a deeply interdisciplinary domain in which engineers actively engaged with philosophers of science and social theorists to tackle the unique complexity and uncertainty of large-scale technological systems whose behavior resisted reduction to simple linear causation (Florman and Vincenti 1992). During this period, engineers questioned the epistemic status of predictive models and debated whether simulations constituted genuine explanations or a simple case of statistical summaries (Oreskes et al. 1994; Winsberg 2019). Then, beginning in the early 1990 s, economic and institutional pressures (e.g., manufacturing, just-in-time supply chains, etc.) began to erode this philosophical engagement as accreditation bodies streamlined engineering curricula to accommodate growing technical electives, often at the expense of humanities and social sciences. As Vermaas et al. (Vermaas et al. 2010) observe, the result is an engineering culture that often views philosophical questions as impractical distractions from real technical work.

From around 2010 onward, ML broadly entered engineering systems, yet this wave largely inherited the reductionist norms (Russell and Norvig 2010). As ML models scaled from thousands to billions of parameters, the persistent assumption remained that statistical accuracy (measured through standard performance metrics) equated to epistemic adequacy. This conflation proved problematic: while statistical accuracy captures predictive correspondence, epistemic adequacy demands richer criteria including causal fidelity (e.g., whether learned associations reflect genuine causal structures rather than spurious correlations), semantic stability (i.e., invariance of representations under meaning-preserving transformations), and explanatory coherence, which shows alignment between model internals and domain-theoretic constructs. The inadequacy of this equivalence became apparent once ML models assumed decision making roles traditionally reserved for human experts (Tulabandhula and Rudin 2014). Rather than developing frameworks for evaluating these epistemic dimensions, industry and regulatory guidance typically reduced associated philosophical problems to procedural checklists (e.g., post hoc explainability tools, fairness metrics). This led to sidestepping deeper questions about the ontology of learned representations or the normative basis of trust (Floridi and Cowls 2022).

However, the philosophical dimensions of ML become apparent when examining how ML systems generate knowledge claims (Floridi 2011; Winfield and Jirotka 2018). On the one hand, traditional engineering relied on deterministic models grounded in physical laws and mathematical proofs. Thus, engineers could trace causal chains through well-understood mechanisms (Popper 2005; Kuhn 2020). On the other hand, ML models operate through statistical correlations. As one can see, the interpretability challenge represents not a technical limitation but a fundamental epistemological shift (Doshi-Velez and Kim 2017). More specifically, engineers increasingly rely on systems whose decision making processes resist human comprehension, yet they often lack the philosophical frameworks necessary to evaluate the epistemic status of systems' outputs (Burrell 2016). Such philosophical disengagement propagates in several observable patterns within engineering organizations, wherein technical discussions frequently treat high-performance metrics as sufficient justification for deployment (O'Neil 2016). As seen in the previous section, several studies note how engineers describe ML systems using anthropomorphic language (without critically examining what these terms mean when applied to statistical models (Agassi and Wiezenbaum 1976). Such linguistic practices reflect deeper conceptual confusions about the nature of intelligence, knowledge, and agency.

The continued rise of ML has also transformed the temporal dynamics of engineering practice. While classical engineering projects operate within well-defined phases (i.e., design, implementation, testing, and maintenance), ML systems continuously evolve through ongoing training. This creates systems whose behavior changes over time in ways that may diverge from initial specifications (Lu et al. 2019). This temporal fluidity challenges traditional engineering

concepts of verification and validation, yet few engineers possess the philosophical tools to articulate why static testing paradigms prove inadequate for dynamic learning systems (Varshney and Alemzadeh 2017).

Despite these mounting challenges, substantial philosophical frameworks for technology design have emerged over the past two decades. For example, the empirical philosophy of technology movement, particularly through value-sensitive design (VSD), has developed sophisticated methodologies for integrating normative considerations into technical systems (Friedman et al. 2015). In particular, Friedman et al.'s (Friedman et al. 2017) multilateral methodology, which encompasses conceptual, empirical, and technical investigations, provides systematic approaches for identifying stakeholder values, translating them into design requirements, and evaluating their implementation in deployed systems. This framework has been successfully adapted to algorithmic systems through what Umbrello and van de Poel (Umbrello and Poel 2021) term *AI for Social Good* to address the unique challenges of autonomous decision making systems. Similarly, van den Hoven et al. (Hoven and P.V.-, and technological design, undefined 2015) demonstrate how design for values approaches can operationalize ethical principles into technical specifications that move beyond post-hoc ethical assessments. However, despite extensive application in human–computer interaction and information systems research, these frameworks remain conspicuously absent from mainstream ML engineering practice, where philosophical engagement typically reduces to compliance with fairness metrics or explainability requirements (Jobin et al. 2019).

This disconnect between available philosophical resources and engineering practice reflects deeper structural impediments to interdisciplinary integration. The *empirical turn* in philosophy of technology, as characterized by Kroes and Meijers (Kroes and Meijers 2001), emphasizes studying actual technological practices rather than abstract theorizing. Yet, ML engineers continue operating with implicit philosophical commitments inherited from statistical learning theory that predate this empirical sophistication. Recent attempts to bridge this gap, such as Selbst et al.'s (Selbst et al. 2019) analysis of abstraction traps in fair ML, reveal how engineers' failure to engage with sociotechnical perspectives leads to systematic oversights in system design. This pattern suggests that the problem extends beyond awareness to encompass fundamental misalignments between engineering's optimization-oriented paradigms and philosophy's critical-reflexive methodologies, requiring not just exposure to philosophical frameworks but transformation of engineering epistemology itself (Winner 2020).

It is worth noting that treatment of epistemological challenges may also arise from institutional constraints and misaligned incentives that prevent them from adequately addressing these concerns. In fact, some of the recent high-profile ML deployments clearly illustrate the consequences of such philosophical blind spots (Schweizer 2022). For example, facial recognition systems deployed without consideration of their epistemological limitations have produced discriminatory outcomes that engineers struggled to anticipate or remediate (Buolamwini and Gebru 2018; Klare et al. 2012). While, in all fairness, some engineers may have likely understood these risks, it is equally possible that organizational pressures for rapid deployment often override philosophical concerns about fairness and representation. These failures stem not solely from inadequate technical skills or insufficient philosophical frameworks, but from a complex interplay of conceptual gaps, institutional constraints, and incentive structures that discourage critical reflection on how ML systems function as sociotechnical artifacts embedded within complex human contexts. The challenge is thus twofold: developing philosophical frameworks for understanding these systems and creating organizational cultures and incentive structures that allow engineers to act on such understanding.

3 A Discussion on Pitfalls Associated with Philosophical Naivety

As noted above, many of the shortcomings at the intersection of ML and engineering emerge not from a lack of mathematical rigor or computational sophistication but from the absence of philosophical frameworks necessary to properly conceptualize, deploy, and interpret ML systems. The following section examines three critical flaws that arise from such a philosophical naivety and demonstrates through engineering examples how abstract conceptual limitations translate into tangible operational failures. Figure 1 shows a visual representation of such flaws and how they interact.

3.1 Ontological Misclassification of System Capabilities

The first and perhaps most pervasive flaw concerns the systematic mischaracterization of what ML systems fundamentally are and what they can meaningfully accomplish (J. Mitchell, R.H.-P. of the 15th A. SIGPLAN-SIGACT, undefined 1988). Here, engineers trained primarily in deterministic systems often project inappropriate ontological categories onto probabilistic ML models. Such categories treat pattern recognition as *understanding* and correlation as *causation* at their core. Such descriptions also suggest intentionality, goals, and comprehension (where only optimization processes exist). Therefore, when engineers

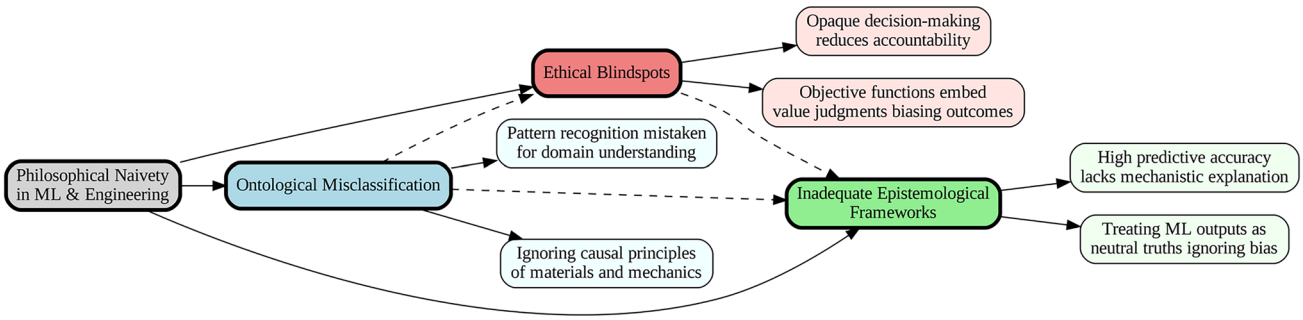
Opaque decision-making reduces accountability
Ethical Blindspots
Objective functions embed value judgments biasing outcomes
Pattern recognition mistaken for domain understanding
Philosophical Naivety in ML & Engineering
Ontological Misclassification
Inadequate Epistemological Frameworks
High predictive accuracy lacks mechanistic explanation
Treating ML outputs as neutral truths ignoring bias
Ignoring causal principles of materials and mechanics

Fig. 1 Pitfalls associated with philosophical naivety

uncritically adopt cognitive terms, they import philosophical notions that obscure crucial differences between human cognition and statistical pattern matching (Wilson 1995; Naser 2025). This misclassification stems from an impoverished philosophical vocabulary that lacks the conceptual tools to distinguish between functional mimicry and comprehension. This can be seen in well-documented studies, which note how anthropomorphic framing can inflate trust and reduce vigilance when system limits are poorly understood. For example, transportation investigations have cited overreliance and misunderstanding of ML system limits as contributing factors in automated driving incidents (Waytz et al. 2014; Ntsb 2017; Parasuraman and Riley 1997). These patterns justify treating capability misclassification as engineering risk controls, not philosophical niceties.

Consider, for example, the deployment of ML systems in the structural health monitoring of bridges. When engineers implement a standard/typical ML model to detect anomalies in vibration patterns, they frequently describe the system as *understanding* structural behavior or *knowing* when damage occurs. This anthropomorphization reflects a deeper ontological confusion between pattern matching and structural comprehension. While a properly trained model might correlate specific frequency signatures with historical failure modes, it possesses no understanding of material science, load distribution, or the physical principles governing structural mechanics. The system operates purely through statistical associations derived from training data, fundamentally different from an engineer's causal understanding grounded in physics and mechanics of materials. Similarly, ML models are often used in mechanical engineering contexts to predict wear and failure in rotating machinery (e.g., gearboxes or turbines). Engineers may claim the model *knows* when a bearing is about to fail, but in reality, it only recognizes patterns in vibration or acoustic data that have historically preceded failure events. Such models may fundamentally lack comprehension of tribology or the complex interactions between surface materials under load.

As one can see, trusting these predictions without understanding the underlying mechanical principles can lead to maintenance decisions that disregard critical factors like operating temperature, potentially causing unexpected breakdowns. Consequently, ontological misclassification becomes particularly hazardous when engineers delegate critical decision making to ML systems (Guizzardi et al. 2023; Hammerschmidt et al. 2025). For instance, engineers who mistake predictive accuracy for genuine understanding may over rely on model outputs during novel operating conditions outside the training distribution. This ontological confusion leads to interventions based on correlational patterns that may not persist when systems alter the underlying causal structure. Such over-reliance may result in failures under certain conditions, a direct consequence of treating statistical inference engines as if they possessed genuine domain knowledge.

3.2 Ethical Blind Spots in ML System Design

The second critical flaw emerges in the ethical dimensions of ML deployment, where philosophical naivety creates systematic blind spots in system design and implementation. More fundamentally, the ethical dimensions of ML systems remain underexamined within engineering practice. As such, engineers often assume that technical optimization can be separated from ethical considerations, treating fairness, justice, and social impact as external constraints rather than integral design parameters (Poel et al. 2011). Specifically, engineers often approach ML development through a purely technical lens that treats ethical considerations as external constraints rather than integral design parameters. This compartmentalization reflects an impoverished understanding of how values become embedded in technical systems through design choices, training data selection, and optimization criteria.

Algorithmic bias provides one of the clearest illustrations of such a blind spot, as engineers frequently approach bias as a purely technical problem (Slota et al. 2020; Esmaeilzadeh 2024). Such a problem is seen to be solvable through dataset balancing or algorithmic adjustments (i.e., without confronting the philosophical questions underlying fairness

metrics: What constitutes fair treatment across different populations? How should historical injustices inform present decisions? Is statistical parity more appropriate than equalized odds for a given context?). In reality, different fairness criteria often prove mathematically incompatible, which forces value trade-offs that cannot be resolved through technical balancing/optimization alone (Hertweck and Räz 2022; Lazar and Stone 2024).

While efficiency metrics serve optimization models well, they obscure distributive questions about who bears residual error. As one might expect, there is a tension between consequentialist frameworks that justify aggregate optimization and deontological approaches that emphasize individual rights. This tension represents a philosophical divide that engineers must navigate explicitly rather than inadvertently resolve through technical choices. Consider a ML wildfire-prediction model whose training objective is to minimize the average spatial deviation between forecast and eventual burn perimeter. Here, by weighting dense suburban areas (where asset counts are high) over sparsely populated tribal lands, the model systematically undervalues risk to marginalized communities. Technically, this occurs when the loss function uses population density or property values as weights (thereby effectively encoding a utilitarian calculus that may conflict with principles of equal protection). This example illustrates how an ostensibly neutral objective function can codify value judgments that disadvantage certain populations.

A similar example can be seen when deploying ML systems for traffic flow optimization. Such a model might optimize for aggregate throughput by reducing average commute times across the network. However, this optimization also embodies hidden value judgments about whose time matters most. The critical distinction lies in whether the system minimizes total travel time across all vehicles (which favors high-traffic routes) vs. minimizing the worst-case travel time for any individual, or ensuring no route exceeds an acceptable threshold. Such a system might systematically route traffic away from affluent neighborhoods and lower-income areas to optimize global efficiency at the cost of local environmental justice.

These examples demonstrate how *ethical reductionism* (i.e., the compression of complex moral considerations into scalar objectives) can prioritize aggregate benefit over individual rights, potentially triggering social backlash. For completion, some possible alternative approaches include: 1) constrained optimization that enforces minimum performance guarantees across subgroups; 2) multi-objective optimization using Pareto frontiers to explicitly visualize efficiency-equity trade-offs; or 3) lexicographic optimization that prioritizes worst-off groups before optimizing aggregate metrics. The choice among these approaches cannot be made on technical grounds alone but requires explicit ethical reasoning and stakeholder engagement.

The problem compounds when ML systems make quasi-judicial decisions, such as automated building code compliance checking or safety inspection prioritization. For example, a ML model trained on historical inspection data might perpetuate past biases that systematically under-inspect certain building types or neighborhoods. Engineers focused solely on prediction accuracy are likely to miss how their systems encode and amplify historical injustices. The absence of philosophical training in moral epistemology leaves them ill-equipped to identify, articulate, or address these embedded biases.

The opacity of ML decision making processes also creates additional ethical challenges that purely technical training cannot address (Lepri et al. 2018). When a ML model recommends structural interventions or safety measures, the inability to provide comprehensible justifications raises profound questions about professional responsibility and public accountability (Laat 2018; Choudhury et al. 2019). Thus, engineers lacking grounding in virtue ethics and deontological frameworks struggle to navigate the tension between leveraging ML tools and maintaining their professional obligation to provide transparent, justifiable recommendations. This implies that even when engineers hold no explicit ethical commitments, modeling choices still embed value judgments. The *abstraction traps* literature shows how failing to treat ML systems as sociotechnical (rather than purely technical) leads to systematic design oversights (Selbst et al. 2019).

3.3 Inadequate Epistemological Frameworks for Model Outputs

The third fundamental flaw concerns the epistemological foundations for interpreting and validating ML-generated knowledge claims (Desai et al. 2022). This epistemological naivety manifests in both overconfidence in model predictions and failure to recognize the implicit philosophical commitments embedded in different modeling approaches. For example, in mechanical engineering applications such as computational fluid dynamics (CFD), engineers increasingly use ML models to approximate complex turbulent flows. While these models might achieve high accuracy on validation sets, they operate through fundamentally different epistemological principles than physics-based simulations. Technically, traditional CFD grounded in Navier–Stokes equations embodies a commitment to mechanistic explanation and causal understanding. On the other hand, ML approximations, conversely, offer predictive power without explanatory insight. This represents a shift from understanding *why* to predicting *what* (Mackenzie 2015). Our

concern is not generic *explainability* but epistemic warrant (e.g., traceable justification chains) to mitigate over-reliance mechanisms observed in automated systems.

This epistemological shift becomes critical when engineers ought to defend design decisions or investigate failures. For example, in engineering settings, relying on opaque models can create a gap between required traceability and what post-hoc explanations can reliably provide (Rudin 2019; Lipton 2018b). More specifically, the black-box nature of ML models can conflict with engineering's epistemic culture that values transparency, traceability, and mechanistic understanding (Geyer et al. 2024). Consider a ML-augmented finite element analysis used in aerospace component design. If such a model predicts stress concentrations leading to design modifications, an engineer might struggle to provide the expected degree of causal explanations traditionally required in such an application (vs. if the same finding were to be generated via a classical finite element model). Moreover, engineers often fail to recognize how different ML architectures embody distinct epistemological commitments (Ng and Jordan 2002). A simple example can be seen when trying to choose between frequentist and Bayesian approaches or between discriminative and generative models.

Similarly, hidden philosophical commitments permeate model interpretation practices. When engineers treat model predictions as revelations about underlying reality rather than artifacts of particular training procedures, they commit to a form of naïve realism[2] that ignores the theory-ladenness of ML outputs. Simply, models trained on biased data reproduce those biases not as unfortunate technical artifacts but as assertions about the world. The epistemological error lies in treating models as neutral observers rather than constructed instruments embodying specific perspectives and limitations.

4 Organizational Liabilities and Erosion of Engineering Authority

The philosophical naivety endemic to contemporary engineering propagates not only as an abstract intellectual deficiency but also as tangible organizational vulnerabilities that compound across multiple operational levels (Boldt 2019; Ratcliff 2013). These vulnerabilities emerge from a fundamental mismatch between the epistemological assumptions underlying traditional engineering practice and the nature of ML systems that increasingly mediate engineering decisions (Rismani et al. 2023; Alvarado 2023). For example, classical risk frameworks in engineering presuppose identifiable failure modes and quantifiable probabilities derived from physical models. However, ML systems confound each assumption since their failure modes emerge from complex interactions between training data biases, architectural choices, and deployment contexts that resist enumeration (Giffen et al. 2022). Thus, ML models' probabilistic outputs reflect not well-characterized physical uncertainties but epistemic limitations whose bounds remain unidentified (Hüllermeier and Waegeman 2021).

This fundamental epistemological mismatch becomes particularly problematic when organizations attempt to apply traditional risk metrics to ML systems (Ratcliff 2013; Funer 2022). For instance, the assignment of discrete probability values to ML failure modes implicitly assumes these systems operate within stable probability spaces (yet distribution shifts can fundamentally alter the underlying probability structure) (Liu et al. 2023). Here, engineers lacking philosophical grounding in the problem of induction fail to recognize that past performance provides limited guarantees about future behavior when the data-generating process itself may change (Berber and Srećković 2024). Similarly, philosophical confusion centers on what constitutes adequate justification for automated decisions. Regulators shaped by deterministic thinking demand explanations that ML systems cannot provide without fundamental architectural compromises that may undermine their performance advantages.

These epistemological challenges compound into organizational dependencies as engineers continue to rely on ML-generated analyses without maintaining an independent capacity to verify these outputs (Nahar et al. 2022). This is likely to develop a dependency through the standard model development steps: first, using ML to accelerate routine calculations, then identifying patterns in complex data, and finally, making design decisions that engineers merely review rather than originate. Each step appears rational in isolation, yet collectively, they produce organizations where critical engineering knowledge exists primarily within inscrutable models rather than human understanding (Ishikawa and Yoshioka 2019). This epistemic displacement of knowledge from human minds to statistical models generates accountability paradoxes that existing organizational structures cannot resolve (Berber and Srećković 2024; Gabison and Xian 2025). When an ML-optimized design fails, responsibility diffuses across the engineers who collected training data, those who selected model architectures, validated performance metrics, and approved deployment (Buiten et al. 2023). However, traditional professional liability frameworks assume individual engineers can explain and justify their technical decisions, but using ML creates liability gaps that legal systems struggle to address

[2] Here used not in the technical sense from the philosophy of perception, but to indicate the uncritical assumption that model outputs transparently mirror reality.

(Liability of Employed Engineers | National Society of Professional Engineers, National Society of Professional Engineers 2023).

These accountability gaps directly challenge the foundational assumptions of engineering ethics codes, which presume professionals can understand the basis for their technical decisions and anticipate their consequences (Hoke 2020). In reality, engineers can rarely fully understand why a ML model recommends particular design parameters, nor can they anticipate how these recommendations might fail in novel contexts (Zafar and Khan 2021). The philosophical naivety about the nature of ML-generated knowledge prevents recognition that these systems produce statistical inferences rather than engineering truths (Funer 2022).

Eventually, this loss of interpretive capacity propagates through subtle degradations in engineering judgment. Engineering teams familiar with ML-generated solutions may lose facility with first-principles reasoning and physical intuition. Specifically, junior engineers trained primarily in ML tools may never develop the complete foundational understanding that enables recognition of when models produce physically implausible outputs. On the other hand, senior engineers may find their expertise devalued as organizations prioritize ML fluency over domain knowledge. The result resembles a form of organizational amnesia where engineering wisdom accumulated over decades dissolves into datasets and model weights.

5 A Technical-Philosophical Framework for Engineers Adopting ML Systems

Because many engineers are taught to use ML like any other software tool, they can inadvertently treat predictive success as if it carried the same epistemic warrant as physics-based models (Anne 2023; P. Wijesinghe, H. Parker, E. Stein, A Case Study 2025). Our contribution is to make the relevant philosophical distinctions explicit and operational—so engineers can test what a model's output actually entitles them to claim, and under which conditions. The resolution of engineering's philosophical naivety requires a systematic framework that integrates philosophical reasoning into ML engineering practice through six interconnected stages. This framework transforms philosophical concepts into engineering methodologies and provides structured approaches for addressing ML systems' epistemological, ontological, and ethical challenges. Unlike traditional engineering frameworks that treat philosophy as an external constraint, this approach positions philosophical reasoning as integral to technical decision making. To operationalize the framework, the engineering organization should treat it as a control plane layered on top of its existing MLOps stack: every model, dataset, and experiment acquires a *philosophical state* alongside the usual lifecycle state, and releases are gated by machine-checkable attestations produced by Stages 1–6. The effect is that philosophical reasoning ceases to be commentary and becomes a deploy-time contract that is as enforceable as a unit test. Figure 2 and Table 1 provide a companion discussion with detailed examples.

5.1 Stage 1: Ontological Classification and Capability Assessment

The framework begins with an ontological classification that replaces imprecise anthropomorphic descriptions with technically accurate capability assessments. This stage addresses the fundamental question: *what does the ML system actually do*, as opposed to *what we metaphorically describe it as doing?* This classification process employs formal methods drawn from the philosophy of mind and cognitive science to distinguish between behavioral competence and cognitive understanding, syntactic pattern matching and semantic comprehension, statistical correlation, and causal reasoning. Operationally, this manifests through automated capability profiling suites that execute compositional stress tests, generate adversarial semantic probes, and compute behavioral consistency scores across systematically perturbed inputs, with results feeding into capability assertion validators that reject deployment containing unsupported anthropomorphic claims.

Consider a ML model trained for medical diagnosis. Engineers commonly describe such systems as learning to recognize diseases or understanding medical images.[3] However,

[3] One might argue that engineers using terms like understands or knows etc. to describe their models as simply being sloppy. We reply that the philosophical naivety we identify does not always concern engineers' technical understanding of their models as statistical inference mechanisms (although this actually occurs especially with engineers who have limited exposure of ML), but rather their insufficient attention to the performative and constitutive effects of technical discourse. In our eyes, when engineers deploy such terms as convenient shorthand, they exhibit *naivety* not only about the computational mechanics of their systems but also about the philosophical implications of linguistic choices in technical communication. In a way, the philosophical naivety lies not in believing their ML models literally *understand*, but in dismissing as inconsequential the ontological commitments embedded in their technical discourse—commitments that propagate through specifications, interfaces, and ultimately shape the sociotechnical systems in which these models operate. Therefore, engineers' expertise in statistical inference does not immunize them from the philosophical consequences of their representational choices, particularly when these choices structure how ML capabilities are operationalized in safety–critical domains. Another notion to keep in mind is that we (as authors/readers) of others' published work have no means of knowing what the authors of such work *imply* when they use specific terms such as *see* or *know* or *understand*. While this may be a simple case of terminology selection, it may also very well *imply* that such authors/researchers *actually* meant the use of that particular

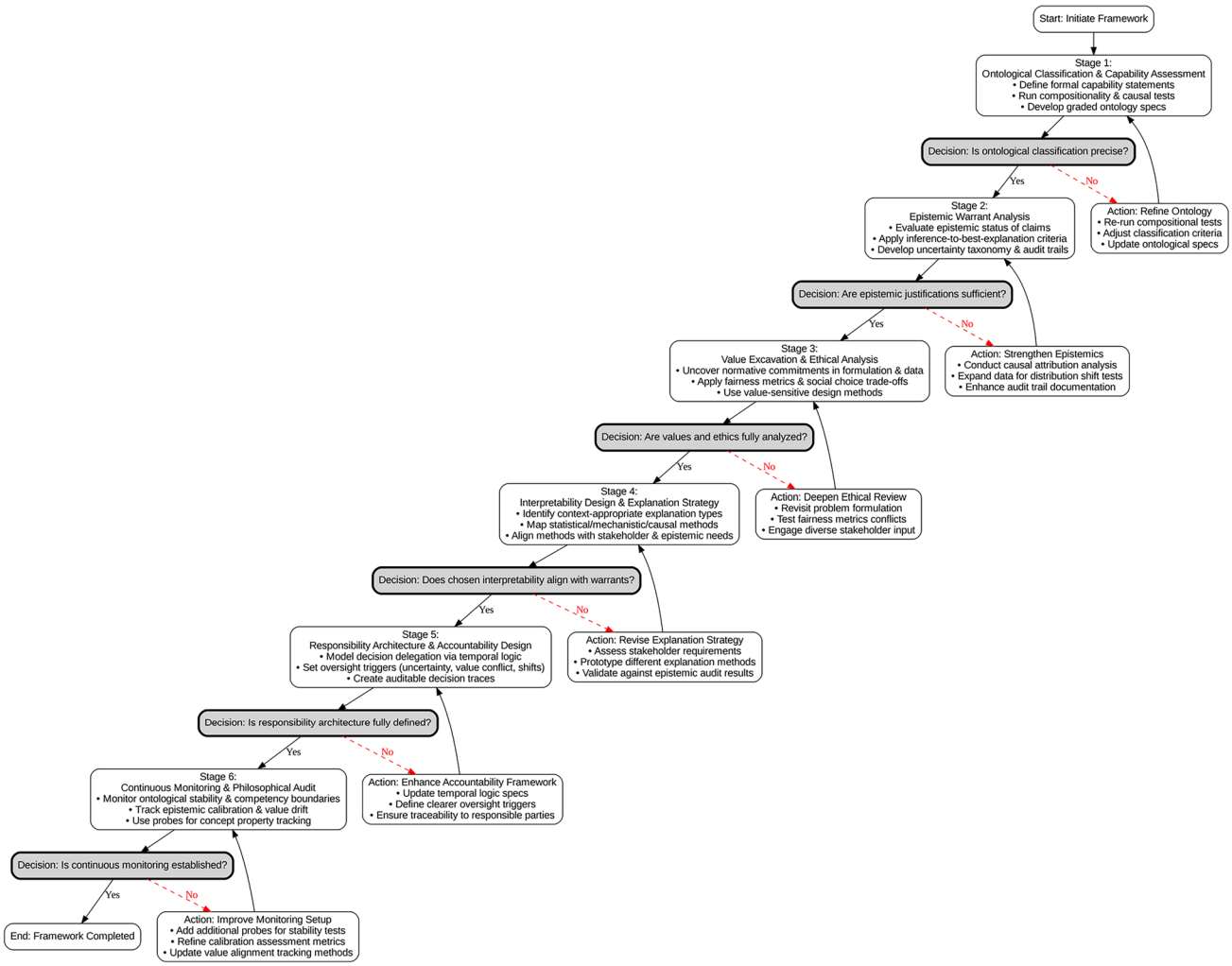
Start: Initiate Framework
Stage 1:
Ontological Classification & Capability Assessment
• Define formal capability statements
• Run compositionality & causal tests
• Develop graded ontology specs
Decision: Is ontological classification precise?
Yes
No
Action: Refine Ontology
• Re-run compositional tests
• Adjust classification criteria
• Update ontological specs
Stage 2:
Epistemic Warrant Analysis
• Evaluate epistemic status of claims
• Apply inference-to-best-explanation criteria
• Develop uncertainty taxonomy & audit trails
Decision: Are epistemic justifications sufficient?
Yes
No
Action: Strengthen Epistemics
• Conduct causal attribution analysis
• Expand data for distribution shift tests
• Enhance audit trail documentation
Stage 3:
Value Excavation & Ethical Analysis
• Uncover normative commitments in formulation & data
• Apply fairness metrics & social choice trade-offs
• Use value-sensitive design methods
Decision: Are values and ethics fully analyzed?
Yes
No
Action: Deepen Ethical Review
• Revisit problem formulation
• Test fairness metrics conflicts
• Engage diverse stakeholder input
Stage 4:
Interpretability Design & Explanation Strategy
• Identify context-appropriate explanation types
• Map statistical/mechanistic/causal methods
• Align methods with stakeholder & epistemic needs
Decision: Does chosen interpretability align with warrants?
Yes
No
Action: Revise Explanation Strategy
• Assess stakeholder requirements
• Prototype different explanation methods
• Validate against epistemic audit results
Stage 5:
Responsibility Architecture & Accountability Design
• Model decision delegation via temporal logic
• Set oversight triggers (uncertainty, value conflict, shifts)
• Create auditable decision traces
Decision: Is responsibility architecture fully defined?
Yes
No
Action: Enhance Accountability Framework
• Update temporal logic specs
• Define clearer oversight triggers
• Ensure traceability to responsible parties
Stage 6:
Continuous Monitoring & Philosophical Audit
• Monitor ontological stability & competency boundaries
• Track epistemic calibration & value drift
• Use probes for concept property tracking
Decision: Is continuous monitoring established?
Yes
No
End: Framework Completed
Action: Improve Monitoring Setup
• Add additional probes for stability tests
• Refine calibration assessment metrics
• Update value alignment tracking methods

Fig. 2 Visual representation of the proposed framework [Note: this figure presents a workflow, not a hierarchy, wherein the questions are equal-priority checkpoints arranged for sequencing and auditability, with permitted iteration and parallelization]

the ontological classification stage requires decomposing these claims into precise technical statements. More specifically, this ML model performs hierarchical feature extraction through convolutional operations that map pixel intensities to probability distributions over diagnostic categories. In a way, this model exhibits behavioral competence in correlating visual patterns with diagnostic labels within its training distribution. However, it lacks a semantic understanding of disease mechanisms, causal reasoning about pathophysiology, or the ability to generate novel diagnostic criteria beyond its training data.

This stage's classification process should utilize formal criteria to allow formality. For example, *compositionality tests* can be employed to examine whether ML systems can recombine learned elements to distinguish true understanding from memorization (Lake and Baroni 2023). These tests operationalize through systematic permutation protocols: generating held-out combinations of atomic features, measuring performance degradation curves, and computing compositionality indices with statistical significance testing via bootstrap confidence intervals. In addition, *causal intervention tests* can help determine the degree of which a ML model operates statistical regularities (e.g., if perturbing semantically irrelevant features (background pixels, image orientation) significantly alters diagnoses, the system relies on spurious correlations rather than meaningful medical understanding) (Jiao et al. 2024). Implementation requires instrumented training pipelines that inject controlled interventions at multiple granularities, semantic object swaps via segmentation masks, and causal graph perturbations using do-calculus operators. Similarly, *counterfactual reasoning assessments* may help evaluate whether systems can answer

term. We believe this is a much more complex item to verify/tackle in this paper, but we are currently considering/it, and we invite interested readers to explore such notion.

Table 1 Additional details and examples related to the proposed framework

Stage	Core purpose	Key questions to address	Technical methods/ tools	Practical examples	Additional considerations
Stage 1: Ontological Classification and Capability Assessment	Replace imprecise anthropomorphic descriptions with technically accurate capability assessments	• What does the ML system do vs. what we metaphorically describe it as doing? • Does the system exhibit behavioral competence or cognitive understanding? • Is it performing syntactic pattern matching or semantic comprehension? • Can it distinguish statistical correlation from causal reasoning? • Does it possess compositional abilities or memorize patterns? • What are the boundaries of its competence domain? • Can it generalize beyond its training meaningfully? • Does it have any form of intentionality or agency? • Can it form abstract concepts or only recognize instances?	• Compositionality tests • Causal intervention tests • Counterfactual reasoning assessments • Behavioral probing experiments • Systematic perturbation analysis • Graded ontology specifications • Formal capability mapping frameworks	**Medical Diagnosis:** Maps pixel intensities to diagnostic probabilities but lacks understanding of disease mechanisms. **Language Models:** Perform syntactic pattern completion without genuine semantic understanding. **Recommendation Systems:** Identify statistical user preferences without understanding human psychology. **Scientific Discovery:** Find correlations in data without understanding theoretical principles	• Document partial capabilities across domains • Establish clear vocabulary for abilities • Create capability matrices with competence gradients • Develop standardized tests for anthropomorphic claims • Map capability boundaries through edge case analysis
Stage 2: Epistemic Warrant Analysis	Evaluate the epistemic status of ML-generated knowledge claims beyond traditional performance metrics	• What types of inferences can this architecture legitimately support? • Does high accuracy imply explanatory power? • Can the model distinguish correlation from causation? • What is the relationship between statistical confidence and epistemic justification? • How do we characterize different types of uncertainty? • When do patterns constitute genuine knowledge discovery vs. curve fitting? • What level of interpretability is required for an epistemic warrant? • How do we validate claimed insights from unsupervised learning? • Can the model support counterfactual reasoning?	• Inference to the best explanation criteria • Causal attribution analysis • Mechanistic interpretability probes • Distribution shift testing • Uncertainty quantification methods • Conformal prediction frameworks • Epistemic audit trails • Formal notation for inference types	**Drug Discovery:** Distinguish between models that learn chemical principles vs. memorize molecular fingerprints. **Climate Models:** Evaluate whether predictions rest on physical understanding or statistical extrapolation. **Economic Forecasting:** Assess if models capture causal mechanisms or historical patterns. **Social Media Analysis:** Determine if sentiment analysis reflects actual understanding or keyword matching. **Genomics:** Evaluate if disease predictions use biological pathways or statistical associations	• Develop uncertainty taxonomies (aleatory vs. epistemic) • Create justification chains from data to decisions • Flag weak epistemic foundations automatically • Establish criteria for different knowledge types • Document assumption dependencies explicitly
Stage 3: Value Excavation and Ethical Analysis	Systematically uncover and evaluate values embedded throughout ML systems	• What values are implicit in the problem formulation? • Whose interests are prioritized in the objective function? • How do data selection choices encode normative commitments? • What ethical theories do our fairness metrics embody? • How do preprocessing decisions affect different populations? • What trade-offs between competing values are we making? • Are we perpetuating historical biases through our training data? • How do we balance individual vs. collective benefits?	• Formal fairness metrics mapping to ethical theories • Value-sensitive design methodologies • Stakeholder impact analysis • Distributional analysis tools • Bias detection algorithms • Ethical review protocols • Participatory design frameworks	**Hiring Algorithms:** Unpack assumptions about 'good employees' and meritocracy. **Healthcare Allocation:** Examine implicit valuations of different health outcomes and populations. **Credit Scoring:** Reveal assumptions about financial responsibility and systemic barriers. **Educational Assessment:** Examine values about intelligence, potential, and standardization. **Criminal Justice:** Uncover assumptions about rehabilitation vs. punishment	• Map technical choices to normative commitments • Document value trade-offs explicitly • Engage diverse stakeholders in value specification • Create value impact assessments • Establish ethical review checkpoints • Develop value alignment metrics

Table 1 (continued)

Stage	Core purpose	Key questions to address	Technical methods/ tools	Practical examples	Additional considerations
Stage 4: Interpretability Design and Explanation Strategy	Develop context-appropriate interpretability mechanisms aligned with epistemic warrant and stakeholder needs	• What type of explanation does this context require? • Who are the stakeholders and what are their explanation needs? • Should explanations be causal, statistical, or mechanistic? • How do we balance accuracy with interpretability? • How do we validate that explanations are faithful? • When is post-hoc explanation sufficient vs. inherent interpretability? • How do we handle explanation uncertainty? • What are the risks of misinterpretation?	• Statistical explanation methods (confidence intervals, prediction sets) • Mechanistic explanation tools (probe classifiers, activation analysis) • Causal explanation frameworks (SCMs, instrumental variables) • Counterfactual generation techniques • Concept activation vectors • Explanation evaluation metrics	**Medical Diagnosis:** Doctors need causal explanations linking findings to diagnoses. **Loan Decisions:** Applicants need actionable explanations for improving approval chances. **Autonomous Vehicles:** Safety engineers need mechanistic understanding of failure modes. **Scientific Discovery:** Researchers need theoretical explanations connecting to domain knowledge. **Legal Applications:** Judges need precedent-based explanations with uncertainty quantification	• Match explanation types to decision contexts • Develop explanation quality metrics • Establish faithfulness criteria • Implement explanation auditing processes
Stage 5: Responsibility Architecture and Accountability Design	Establish mechanisms for maintaining human responsibility despite algorithmic mediation	• How do we maintain meaningful human control? • What decisions require human approval vs. full automation? • How do we create traceable accountability chains? • When should systems refuse to make decisions? • How do we handle responsibility for emergent behaviors? • What oversight mechanisms are proportional to impact? • How do we prevent responsibility diffusion? • What documentation enables post-hoc analysis?	• Statistical process control for ML • Oversight trigger mechanisms • Accountability trace protocols • Decision delegation frameworks • Responsibility assignment matrices • Human-in-the-loop architectures • Fallback procedure design	**Healthcare:** Establish when ML recommendations require physician review based on certainty and severity. **Financial Trading:** Create circuit breakers for anomalous AI behavior with human override. **Content Autonomous Systems:** Implement handoff protocols when exceeding operational boundaries. **HR Systems:** Require human validation for high-stakes employment decisions. **Legal Tech:** Maintain attorney responsibility for AI-assisted document review	• Define clear responsibility boundaries • Establish liability frameworks • Design for appropriate human oversight • Implement continuous monitoring systems • Develop incident response protocols
Stage 6: Continuous Monitoring and Philosophical Audit	Establish ongoing evaluation of philosophical assumptions as systems evolve	• Are systems maintaining consistent capability boundaries? • Is there evidence of competence creep beyond training domains? • Do confidence estimates remain well-calibrated over time? • Are embedded values drifting from stakeholder intentions? • How do we detect emergent behaviors or capabilities? • When do systems require re-evaluation of ontological status? • What triggers fundamental reassessment? • How do we track anthropomorphic misunderstanding?	• Ontological stability tests • Epistemic calibration tracking • Value drift detection algorithms • Conceptual property metrics • Behavioral consistency monitoring • Stakeholder alignment surveys • Emergent behavior detection • Philosophical assumption validators	**Recommendation Systems:** Track value drift as user preferences evolve and feedback loops form. **Healthcare:** Detect when systems exceed competence in new disease areas. **Social Platforms:** Track how content curation values shift with user engagement patterns. **Autonomous Agents:** Detect emergent goal-seeking behaviors beyond original programming	• Implement continuous philosophical monitoring • Create reassessment trigger criteria • Document assumption evolution • Establish regular audit cycles • Design for system introspection • Maintain alignment with evolving values

what if questions that require causal models rather than pattern matching (Verma et al. 2024).

The technical implementation of this stage involves constructing formal capability matrices using description logic that encodes subsumption hierarchies of competencies to enable automated reasoning about capability inheritance and conflict detection when multiple models compose into ensemble systems. In addition, such implementations can benefit from ontological specification languages that capture these distinctions formally. Rather than binary

classifications, the framework prefers to employ graded ontologies that acknowledge that ML systems may exhibit partial properties. For instance, a language model might demonstrate syntactic competence without semantic understanding or possess local semantic capabilities for specific domains without general comprehension. These specifications feed into subsequent stages to ensure that deployment decisions align with actual rather than imagined capabilities. The pipeline integration occurs through capability manifest files (e.g., YAML/JSON schemas) that gate production deployments. In such an integration, models failing to meet minimum capability thresholds for their declared use cases trigger automated rollback with detailed deficit reports linking failed tests to missing ontological properties.

5.2 Stage 2: Epistemic Warrant Analysis

The second stage develops formal methods for evaluating the epistemic status of ML-generated knowledge claims. This analysis goes beyond traditional performance metrics to examine what kinds of inferences different ML architectures can legitimately support. The challenge involves bridging the gap between statistical confidence and epistemic justification, with a clear reminder that high predictive accuracy does not automatically confer explanatory power or causal understanding. The analysis within this stage can also leverage technical tools from formal epistemology and philosophy of science (Hendricks 2005). For example, *inference to the best explanation* criteria can help determine when pattern recognition constitutes genuine knowledge discovery versus simple curve fitting (Harman 1965). This stage may also adapt philosophical accounts of explanation (e.g., deductive-nomological, causal-mechanical, unificationist, etc.) into computational criteria for evaluating ML outputs (Kitcher and Salmon 1962; Hempel and Oppenheim 1948).

This stage introduces uncertainty taxonomies to distinguish aleatory uncertainty (inherent randomness in data) from epistemic uncertainty (limitations in model knowledge). In parallel, the framework supports conformal prediction methods to quantify predictive uncertainty, but one should note that epistemic warrant analysis goes further to examine the sources and implications of this uncertainty (Shafer and Vovk 2008). This can be beneficial, especially when deploying ML for critical decisions, as engineers must determine whether uncertainty stems from data noise (potentially reducible through more samples) or fundamental limitations in the modeling approach (requiring architectural changes or hybrid methods).

Operationalization at this stage requires constructing epistemic certifications that could compute warrant scores through multi-criteria decision analysis by combining statistical reliability metrics (e.g., calibration error, Brier scores), causal validity indicators (e.g., instrumental variable strength, backdoor adjustment completeness), and mechanistic interpretability measures (e.g., feature attribution stability) into composite epistemic health indices. Each explanatory can translate into an executable validator, such as deductive-nomological checkers verify learned rules against domain axioms, causal-mechanical analyzers trace prediction paths through structural equation models, and unificationist scorers compute compression ratios of phenomenon descriptions given learned representations. Here, a prediction backed by learned causal structure warrants stronger epistemic claims than one based solely on correlational patterns (Woodward 2004). For instance, when a graph neural network trained on molecular data predicts drug efficacy, epistemic warrant analysis examines the basis for this prediction. Does the network learn chemically meaningful features corresponding to functional groups and binding sites? Can it generalize to novel molecular scaffolds through compositional reasoning? Or does it memorize statistical patterns from the training set?

In practical terms, the implementation of this stage involves developing epistemic audit trails that document the justification chain from raw data through model predictions to decision recommendations. Audit trails materialize as directed acyclic graphs (DAGs) in provenance tracking systems with nodes annotated by epistemic metadata: confidence degradation factors, assumption violation flags, and inferential leap indicators computed through symbolic execution of the prediction pipeline. Such trails may employ formal notation systems to capture inference types, assumption dependencies, and confidence degradation through the pipeline. The same trails can be automated to flag instances where strong claims rest on weak epistemic foundations, such as causal conclusions drawn from observational data without appropriate identification strategies (O'Hara 2022). Therefore, flagging triggers when identification assumptions fail validation tests, with severity scores computed from sensitivity analysis (e.g., how much must assumptions be violated before conclusions reverse?).

5.3 Stage 3: Value Excavation and Ethical Analysis

The proposed framework recognizes that technical choices inherently encode normative commitments; thus, the third stage systematically uncovers and evaluates values embedded throughout ML systems (Jobin et al. 2019). This process moves beyond surface-level bias detection to examine how problem formulations, data selection, feature engineering, and optimization objectives collectively constitute ethical frameworks. Value excavation begins with problem formulation analysis. When engineers define an ML task, they

make normative judgments about what outcomes matter, whose interests count, and what constitutes success.

For example, a hiring algorithm optimizing for *employee quality* embeds values about what makes a good employee, whose definition of quality prevails, and whether past hiring patterns should guide future decisions. Therefore, the framework provides structured methods for decomposing these high-level objectives into constituent value commitments and examining their implications. Some of the companion tools that can be used herein include formal fairness metrics that capture different normative theories of justice (e.g., individual fairness requires treating similar individuals similarly, while group fairness demands statistical parity across protected categories) (Oneto et al. 2024; Zhao et al. 2023).

The framework demonstrates how these criteria often conflict mathematically, which leads to explicit value tradeoffs. As impossibility theorems from social choice theory prove that no single metric can satisfy all reasonable fairness desiderata simultaneously, then engineers are likely to engage in substantive ethical reasoning about which values to prioritize in specific contexts (Bell et al. 2023). Some scenarios that engineers might face in this stage include those related to dataset construction (and the numerous decisions about inclusion, representation, and categorization that shape model behavior), feature engineering/preprocessing (as they encode values through decisions about relevant distinctions and categories), as well as questions such as how does binning continuous variables affect different populations? What values are implicit in choosing certain distance metrics or kernel functions? How do normalization schemes differentially impact various groups? It is worth noting that this stage could employ technical methods from value-sensitive design, feminist epistemology, and critical algorithm studies to make explicit the values that traditional engineering practice leaves implicit (Sadek et al. 2023).

5.4 Stage 4: Interpretability Design and Explanation Strategy

The fourth stage develops context-appropriate interpretability mechanisms that align with epistemic warrant levels and stakeholder needs. Rather than pursuing universal interpretability, the framework recognizes that different contexts demand different explanation types (Glennan 2002; Grace and Irvine 2020). Context mapping often employs decision criticality matrices crossing consequence severity (e.g., reversibility, magnitude, scope) with decision frequency to generate explanation requirement specifications that mandate explanation types, granularities, and validation standards for each cell, with automated matching of ML outputs to appropriate explanation generators. Therefore, this stage introduces a taxonomy of explanation types mapped to technical methods. Some types may include *statistical* explanations that employ confidence intervals, prediction sets, and influence functions to quantify uncertainty and identify critical training examples, as well as *mechanistic* explanations that use probe classifiers and activation analysis to reveal internal representations, and *causal* explanations that leverage structural causal models, instrumental variables, counterfactuals, and mediation analysis to identify intervention points. As one might expect, this stage also provides criteria for selecting appropriate explanation methods: does the context require understanding why a prediction was made (requiring causal or mechanistic explanation) or how confident we should be (allowing statistical explanation)?

5.5 Stage 5: Responsibility Architecture and Accountability Design

This stage of the proposed framework addresses the challenge of maintaining professional responsibility when engineers cannot fully understand or predict ML system behavior (Cooper et al. 2022; Wieringa 2020). Thus, the fifth stage establishes mechanisms for maintaining human responsibility despite algorithmic mediation of decisions. This involves technical methods for preserving meaningful human control that allow appropriate oversight capabilities and create accountability traces from automated decisions back to responsible parties (Diakopoulos 2016). One can think of this stage as developing a *responsibility architecture* that begins with a formal analysis of decision delegation patterns. The framework employs temporal logic specifications to model decision flows: which decisions can be fully automated, require human approval, and require human initiation with ML support? These specifications account for context-dependent factors, including decision reversibility, impact magnitude, and available human expertise.

This stage also provides criteria for setting oversight triggers based on epistemic uncertainty, value conflicts, or distribution shift detection. These triggers employ statistical process control methods adapted for ML contexts to distinguish systematic degradation from random variation. In addition, the proposed framework can be augmented with accountability protocols that create auditable traces from ML recommendations through human decisions to outcomes (especially when adverse outcomes occur, such traces can enable a quick and systematic analysis of whether failures stemmed from ML limitations, human oversight failures, or unavoidable uncertainty).

5.6 Stage 6: Continuous Monitoring and Philosophical Audit

The final stage establishes mechanisms for the ongoing evaluation of philosophical assumptions as ML systems evolve through continued training and emerging use cases (Dignum 2019). This philosophical monitoring employs technical probes targeting specific conceptual properties. Probes may implement containerized microservices with standardized APIs, executing philosophical health checks on configurable schedules, publishing metrics to databases with anomaly detection, and triggering PagerDuty alerts when philosophical drift exceeds acceptable bounds. For instance, monitoring may cover several fronts, namely *ontological stability tests* (to examine whether systems maintain consistent capability boundaries or exhibit competence creep into domains beyond their training), *epistemic calibration assessments* (to track whether confidence estimates remain reliable as data distributions shift), and *value drift detection* (to analyze whether embedded values remain aligned with stakeholder intentions as systems adapt to new data). Such monitoring aims to understand failures: did anthropomorphic misunderstanding of capabilities contribute to inappropriate deployment? Did unexamined value assumptions lead to discriminatory outcomes? Did epistemic overconfidence result in decisions beyond the system's warrant? Practically, engineers can track metrics related to conceptual properties (ontological confusion rates in system descriptions), epistemic calibration scores across decision types, and value alignment measures relative to stated principles.

5.7 Metrics of Success and Long-Term Vision

The evaluation of philosophical integration within engineering practice requires careful consideration of quantitative and qualitative metrics that capture the multidimensional nature of epistemic improvement (Langford 2016; Mari et al. 2023). While valuable for assessing technical performance, traditional engineering metrics (e.g., quantitative indicators to assess technical performance, efficiency, and reliability) can prove insufficient when measuring the success of philosophical engagement in ML-mediated engineering contexts (Goldman 2004). Therefore, the development of appropriate success metrics necessitates a reconceptualization of what constitutes engineering merit in an era where epistemological sophistication directly impacts reliability and ethical soundness (Schuette 1999; Zwart 2022).

From this lens, *short-term* success metrics can be tied to capturing immediate improvements in engineering practice resulting from enhanced philosophical literacy. These include measurable reductions in ontological misclassification rates, where engineers demonstrate improved accuracy in categorizing ML system capabilities and limitations (Zemmouchi-Ghomari 2023). Such improvements can be gauged through decreased instances of anthropomorphic attribution in technical documentation, more precise delineation between correlation and causation in model interpretations, and enhanced recognition of the boundaries between deterministic and probabilistic system behaviors (Salles et al. 2020). Additionally, the frequency and sophistication of ethical considerations documented during design reviews provide quantifiable indicators of philosophical integration (similar to explicit epistemological justifications in technical decision making processes) (Dias 2019).

Medium-term metrics encompass organizational transformations (Grimson and Murphy 2015). For instance, the establishment of cross-functional teams incorporating philosophers of science alongside engineers represents a structural metric of institutional commitment to philosophical integration (Ewim et al. 2024). The quality and frequency of interdisciplinary dialogue, measured through collaborative publication rates and joint project outcomes, can also be used to gauge the depth of integration achieved (Naser 2024; McCance and Blanchard 2024). Furthermore, the evolution of engineering curricula to include mandatory philosophy of science components and the development of hybrid technical-philosophical competency frameworks provide concrete evidence of systemic change (Lysak 2024; Skorton et al. 2018). Such an educational component is likely to produce engineers capable of navigating the complex philosophical terrain inherent in creating systems that increasingly approximate cognitive functions previously exclusive to human intelligence (Wilson 1995; Boon 2011). The same may also lead to the emergence of new professional roles (e.g., epistemological auditors) for ML systems, which can signal the institutionalization of philosophical thinking within engineering organizations.

Long-term success can be demonstrated through fundamental shifts in engineering culture and practice where professional engineering organizations continue to evolve to support this expanded conception of engineering competence (Wilson 1995; Morison and Florman 1989). For example, the development of new engineering methodologies that inherently incorporate philosophical reflection represents an exemplary transformation beyond additive improvements (Schön 2017). These methodologies would treat epistemological analysis as integral to system design rather than as post-hoc validation (Brey 2010). A solid example could revolve around a notion mentioned earlier by creating industry-wide standards for philosophical rigor in ML system development (i.e., analogous to current safety and reliability standards) (ISO 2023). Most significantly, the emergence of a new generation of engineers who naturally synthesize technical and philosophical thinking would mark

the successful cultural transformation necessary for sustainable change (Goldberg 2010; Hyldgaard Christensen 2015). This may lead to a new synthesis where technical and philosophical thinking become indistinguishable aspects of a unified engineering methodology (Boon and Knuuttila 2009). In the long run, such integration can enable engineers to create ML systems that not only perform their intended functions but do so in ways that align with human values and epistemological integrity (Hemphill 2020).

6 Future Research Directions for Industry, Academia, and Professional Societies

The transformation of engineering practice through philosophical integration requires coordinated and interdisciplinary research efforts (Boon and Van Baalen 2019). Such efforts are expected to tie in the complexity of ML systems and their epistemological implications in navigating the boundaries between technical (engineering) and philosophical inquiry (Alvarado 2023; Trans et al. 2024). Here, we discuss possible research directions and implementation strategies for industry, academia, and professional societies to address the philosophical naivety currently limiting engineering practice in the ML era. It is worth noting that this section also consolidates some practices that are already emerging across settings. We aim to showcase and extend such practices with engineering-specific contributions.

For a start, academic institutions can foster foundational research at the intersection of the philosophy of science and engineering (Žukauskas et al. 2018). A possible research direction could include the development of formal frameworks for epistemological analysis of ML systems by incorporating insights from the philosophy of mind to better characterize the nature of machine *understanding* and *reasoning* (Bai 2022). These frameworks must move beyond metaphorical applications of cognitive concepts to establish thorough criteria for evaluating the epistemological status of ML model outputs (Doroudi 2024; Jabareen and Y.J.-I. journal of qualitative methods, undefined 2009). For instance, future research could investigate how different philosophical positions on consciousness and intelligence translate into tangible engineering design decisions (Leavitt et al. 2024). This may allow the creation of various mappings between philosophical commitments and technical architectures (Blackman et al. 2017).

Other research directions can revolve around establishing dedicated research centers for the philosophy of engineering. Such centers would foster collaboration between philosophers (specializing in epistemology, ethics, and philosophy of mind) and engineers (working on creating or applying ML systems) (Newman 2024). Possible research priorities may include developing new ontological categorization schemes for ML capabilities that avoid both anthropomorphic overstatement and reductive mechanistic descriptions, investigating the epistemological implications of different ML architectures, examining how choices in model design embed implicit philosophical assumptions about knowledge, reasoning, and intelligence, to name a few. A particular research direction with a more direct linkage to education may investigate pedagogical approaches for integrating philosophical content into engineering curricula (Delahunty and Kimbell 2021; Matta 2022).

In parallel, industry research must focus on translating philosophical insights into practical engineering methodologies (Zwart 2022; Vries 2021). For instance, corporate research divisions may establish partnerships with academics to develop frameworks for philosophical analysis in commercial ML development (e.g., creating automated tools for detecting ontological misclassifications in system specifications, developing metrics for quantifying epistemological uncertainty in model predictions, establishing ethical review protocols incorporating sophisticated philosophical analysis, etc.) (Rybnicek and Königsgruber 2019). Along the same lines, industry researchers can also investigate how philosophical sophistication impacts system reliability and safety, as well as examine the epistemological implications of using ML systems to evaluate other ML systems. As one might imagine, such collaborations can generate open-source datasets and tools for philosophical analysis of ML systems, including libraries for ontological classification, epistemological uncertainty quantification, and ethical impact assessment – all of which are likely to enable systematic evaluation of progress in philosophical integration (Ahmed et al. 2022).

Additional research directions may include longitudinal studies tracking philosophical integration's impact on engineering outcomes (Rossoni et al. 2024). Such studies can establish cohorts of philosophically-trained engineers to compare their performance against traditional engineering populations across multiple dimensions (e.g., professional identity formation, etc.) (Michelfelder et al. 2016; Mody 2021). These studies may also trace how philosophical sophistication influences career trajectories, team dynamics, and organizational culture within engineering contexts (Islam et al. 2018). The design of these studies can investigate sociological factors that contribute to philosophical naivety among engineers (Vallor 2020). The outcome of these longitudinal studies can develop targeted interventions that facilitate cultural change within the engineering domain (Tereshchenko et al. 2024). Overall, the collection of comprehensive data on the practical impacts of philosophical integration can be crucial for motivating widespread

adoption and gauging its effectiveness across various engineering domains.

Finally, while our analysis deliberately maintains a general treatment of ML, we acknowledge that the core epistemic assumption of equating statistical metrics with epistemic adequacy cuts across most ML variants and that the reductionist inheritance from statistical learning theory operates field-wide rather than variant-specific, and that the proceduralization of philosophical concerns into checklists represents a systemic pattern surpassing specific ML architectures. Therefore, we invite future work to examine how such generalizations of epistemic inadequacies affect ML variants. Such investigations might explore hallucination in LLMs as a particular failure mode of conflating token probability with truth, shortcut learning in vision models as revealing the poverty of accuracy-based evaluation, or the causal opacity of ensemble methods as challenging interpretability frameworks designed for single models. Each variant's unique inductive biases and optimization dynamics likely generate distinct philosophical challenges that, while rooted in the systemic issues we identify here, deserve dedicated analysis to fully understand their epistemic implications.

7 Conclusions

Engineering's widespread adoption of ML has exposed a critical philosophical naivety, manifesting as systematic misunderstandings about ML's fundamental nature and limitations. This distinction is often missed and continues to lead to ontological misclassification, where engineers mistakenly attribute cognitive capabilities like understanding to systems performing pattern matching. Furthermore, ethical blind spots arise from treating optimization objectives as value-neutral, which obscures significant distributive consequences, while inadequate epistemological frameworks result in brittle validation and acceptance of spurious correlations. These issues undermine engineering's epistemic authority and challenge traditional risk management. Addressing these liabilities requires a fundamental integration of philosophical reasoning into engineering practice, as outlined in the proposed technical-philosophical framework, supported by coordinated efforts across industry, academia, and professional societies, and measured by evolving metrics that value philosophical literacy alongside technical skill. Additional key findings include:

- Engineers may misclassify ML output by treating statistical patterns as if they possess causal insight comparable to mechanics-based simulators.
- ML prediction often lacks the explanatory transparency inherent in traditional engineering models, challenging the need to understand why.
- Applying traditional risk frameworks to ML is difficult due to complex, non-enumerated failure modes arising from data and architecture interactions.
- Heavy reliance on ML can lead to a degradation of engineering judgment, as intuition and first-principles reasoning may decline.
- Addressing philosophical naivety requires coordinated research and cultural shifts across academia, industry, and professional societies.

Funding Open access funding provided by the Carolinas Consortium. None

Data Availability Data is available on request from the author.

Declarations

Conflict of interest The author declares no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Agassi J, Wiezenbaum J (1976) Computer power and human reason: from judgment to calculation. Technol Cult. https://doi.org/10.2307/3103715

Ahmed F, Fattani MT, Ali SR, Enam RN (2022) Strengthening the bridge between academic and the industry through the academia-industry collaboration plan design model. Front Psychol. https://doi.org/10.3389/fpsyg.2022.875940

Alvarado R (2023) Ai as an epistemic technology. Sci Eng Ethics. https://doi.org/10.1007/s11948-023-00451-3

Andrews RW, Lilly JM, Srivastava D, Feigh KM (2023) The role of shared mental models in human-AI teams: a theoretical review. Theor Issues Ergon Sci. https://doi.org/10.1080/1463922X.2022.2061080

Anne M (2023) How can engineering education embrace ChatGPT? | Colorado State University, Colorado State University. https://engr.source.colostate.edu/how-can-engineering-education-embrace-chatgpt-by-treating-it-like-a-tool/?utm_source=chatgpt.com (accessed October 3, 2025).

Bai H (2022) The Epistemology of Machine Learning, Filosofija. Sociologija. https://doi.org/10.6001/fil-soc.v33i1.4668

Barocas S, Hardt M, Narayanan A (2023) Fairness and machine learning: limitations and opportunities

Bell A, Bynum L, Drushchak N, Zakharchenko T, Rosenblatt L, Stoyanovich J (2023) The possibility of fairness: revisiting the impossibility theorem in practice. ACM Int Conf Proceeding Ser. https://doi.org/10.1145/3593013.3594007

Berber A, Srećković S (2024) When something goes wrong: who is responsible for errors in ML decision-making? AI Soc. https://doi.org/10.1007/s00146-023-01640-1

Blackman D, K Moon (2017) A guide to ontology, epistemology, and philosophical perspectives for interdisciplinary researchers, Integration and Implementation Insights.

Boldt J (2019) The concept of vulnerability in medical ethics and philosophy. Philos Ethics Humanit Med. https://doi.org/10.1186/s13010-019-0075-6

Boon M (2011) In defense of engineering sciences: on the epistemological relations between science and technology. Techne Res Philos Technol. https://doi.org/10.5840/techne20111515

Boon M, Knuuttila T (2009) Models as epistemic tools in engineering sciences. Philos Technol Eng Sci. https://doi.org/10.1016/B978-0-444-51667-1.50030-6

Boon M, Van Baalen S (2019) Epistemology for interdisciplinary research – shifting philosophical paradigms of science. Eur J Philos Sci. https://doi.org/10.1007/s13194-018-0242-4

Brey P (2010) Philosophy of technology after the empirical turn. Techné: Research in philosophy and technology. https://doi.org/10.5840/techne20101416

Buiten M, De Streel A, MP-CL & S Review, undefined (2023) The law and economics of AI liability, Elsevier. (n. d.). https://www.sciencedirect.com/science/article/pii/S0267364923000055. Accessed 4 June 2025.

Buolamwini J, Gebru T (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, in: Proc Mach Learn Res

Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data Soc. https://doi.org/10.1177/2053951715622512

Choudhury S, Lee M, Kurenkov A (2019) In Favor of Developing Ethical Best Practices in AI Research, The Stanford AI Lab Blog.

Christie M, de Graaff E (2017) The philosophical and pedagogical underpinnings of active learning in engineering education. Eur J Eng Educ. https://doi.org/10.1080/03043797.2016.1254160

Cohen JP, Cao T, Viviano JD, Huang CW, Fralick M, Ghassemi M, Mamdani M, Greiner R, Bengio Y (2021) Problems in the deployment of machine-learned models in health care. CMAJ 193:E1391–E1394. https://doi.org/10.1503/CMAJ.202066/TAB-RELATED-CONTENT

Cooper AF, Moss E, Laufer B, Nissenbaum H (2022) Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. ACM Int Conf Proceeding Ser. https://doi.org/10.1145/3531146.3533150

de Laat PB (2018) Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? Philos Technol. https://doi.org/10.1007/s13347-017-0293-z

De Vries MJ (2021) Innovation Research in Technology and Engineering Management: A Philosophical Approach. https://doi.org/10.4324/9781003013556

Delahunty T, Kimbell R (2021) (Re)framing a philosophical and epistemological framework for teaching and learning in STEM: emerging pedagogies for complexity. Br Educ Res J. https://doi.org/10.1002/berj.3706

Desai J, Watson D, Wang V, Taddeo M, Floridi L (2022) The epistemological foundations of data science: a critical review. Synthese. https://doi.org/10.1007/s11229-022-03933-2

Diakopoulos N (2016) Accountability in algorithmic decision making. Commun ACM. https://doi.org/10.1145/2844110

Dias P (2019) Philosophy for Engineering. https://doi.org/10.1007/978-981-15-1271-1

Dignum V (2019) Responsible artificial intelligence: how to develop and use AI in a responsible way, Springer International Publishing. Cham. https://doi.org/10.1007/978-3-030-30371-6

Doroudi S (2024) On the paradigms of learning analytics: Machine learning meets epistemology. Computers and Education: Artificial Intelligence 6:100192. https://doi.org/10.1016/j.caeai.2023.100192

Doshi-Velez F, B Kim B (2017) A Roadmap for a Rigorous Science of Interpretability

Esmaeilzadeh P (2024) Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: a perspective for healthcare organizations. Artif Intell Med 151:102861. https://doi.org/10.1016/J.ARTMED.2024.102861

Ewim CP-M, Ewim CP-M, Achumie GO, Adeleke AG, Okeke IC, Mokogwu C (2024) Developing a cross-functional team coordination framework: A model for optimizing business operations, International Journal of Frontline Research in Multidisciplinary. Studies 4:015–034. https://doi.org/10.56355/IJFRMS.2024.4.1.0030

Floridi L (2011) The Philosophy of Information https://doi.org/10.1093/acprof:oso/9780199232383.001.0001

Floridi L, Cowls J (2022) A unified framework of five principles for AI in society, In: Mach Learn City Appl Archit Urban Des https://doi.org/10.1002/9781119815075.ch45

Florman SC, Vincenti WG (1992) What engineers know and how they know it: Analytical studies from aeronautical history. Technol Cult. https://doi.org/10.2307/3105813

Friedman B, Kahn PH, Borning A (2015) Value senstive design and information systems. Human-Computer Interact Manag Inf Syst Found. https://doi.org/10.4324/9781315703619-27

Friedman B, Hendry DG, Borning A (2017) A survey of value sensitive design methods. Foundations and Trends® in Human–computer Interaction. https://doi.org/10.1561/1100000015

Funer F (2022) Accuracy and interpretability: struggling with the epistemic foundations of machine learning-generated medical information and their practical implications for the doctor-patient relationship. Philos Technol. https://doi.org/10.1007/s13347-022-00505-7

Furey H, Hill S, Bhatia SK (2021) Beyond the code: a philosophical guide to engineering ethics. Taylor and Francis. https://doi.org/10.4324/9781315643816/BEYOND-CODE-HEIDI-FUREY-SUJATA-BHATIA-SCOTT-HILL

Gabison G, Xian RP (2025) Inherent and emergent liability issues in LLM-based agentic systems: a principal-agent perspective, Queen Mary Law Research Paper No. 447/2025. https://doi.org/10.2139/SSRN.5210666.

Geyer P, Singh MM, Chen X (2024) Explainable AI for engineering design: a unified approach of systems engineering and component-based deep learning demonstrated by energy-efficient building design. Adv Eng Inform 62:102843. https://doi.org/10.1016/J.AEI.2024.102843

Glennan S (2002) Rethinking mechanistic explanation. Philos Sci. https://doi.org/10.1086/341857

Goldberg DE (2010) The missing basics and other philosophical reflections for the transformation of engineering education. Holist Eng Educ beyond Technol. https://doi.org/10.1007/978-1-4419-1393-7_13

Goldberg DE, Somerville M (2015) The making of a whole new engineer: four unexpected lessons for engineering educators and education researchers. J Eng Educ. https://doi.org/10.1002/jee.20064

Goldman SL (2004) Why we need a philosophy of engineering: a work in progress. Interdiscip Sci Rev. https://doi.org/10.1179/030801804225012572

Grace JB, Irvine KM (2020) Scientist's guide to developing explanatory statistical models using causal analysis principles. Ecology. https://doi.org/10.1002/ecy.2962

Grimson W, Murphy M (2015) The Epistemological Basis of Engineering, and Its Reflection in the Modern Engineering Curriculum. Philos Eng Technol. https://doi.org/10.1007/978-3-319-16172-3_9

Guizzardi R, Amaral G, Guizzardi G, Mylopoulos J (2023) An ontology-based approach to engineering ethicality requirements. Softw Syst Model. https://doi.org/10.1007/s10270-023-01115-3

Hammerschmidt T, Hafner A, Stolz K, Passlack N, Posegga O, Gerholz KH (2025) A Review of How Different Views on Ethics Shape Perceptions of Morality and Responsibility within AI Transformation. Springer. https://doi.org/10.1007/S10796-025-10596-0

Harman GH (1965) The inference to the best explanation. Philos Rev. https://doi.org/10.2307/2183532

Hempel CG, Oppenheim P (1948) Studies in the logic of explanation. Philos Sci 15:135–175. https://doi.org/10.1086/286983

Hemphill TA (2020) Human Compatible: Artificial Intelligence and the Problem of Control, Cato Journal.

Hendricks VF (2005) Mainstream and Formal Epistemology. https://doi.org/10.1017/CBO9780511616150

Hertweck C, Räz T (2022) Gradual (In)Compatibility of Fairness Criteria, In: Proc 36th AAAI Conf Artif Intell AAAI 2022. https://doi.org/10.1609/aaai.v36i11.21450.

Hoke T (2020) ASCE Adopts New Code of Ethics. Civil Engineering Magazine. https://doi.org/10.1061/ciegag.0001546

Van den Hoven J, PV (2015) and technological design, undefined 2015, Design for values: An introduction, Springer. pp. 1–7 https://doi.org/10.1007/978-94-007-6970-0_40.

Hüllermeier E, Waegeman W (2021) Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn. https://doi.org/10.1007/s10994-021-05946-3

Hyldgaard Christensen S (2015) Engineering Identities, Epistemologies and Values.

Ishikawa F, Yoshioka N (2019) How Do Engineers Perceive Difficulties in Engineering of Machine-Learning Systems? - Questionnaire Survey, In: Proc - 2019 IEEE/ACM Jt 7th Int Work Conduct Empir Stud Ind 6th Int Work Softw Eng Res Ind Pract CESSER-IP 2019. https://doi.org/10.1109/CESSER-IP.2019.00009.

Islam S, Lahti K, Chetta MH (2018) Research partnerships between academics and consulting firms: a stakeholder analysis. Ind Organ Psychol. https://doi.org/10.1017/iop.2018.121

ISO, ISO/IEC 23053:2022 (2023) Framework for AI systems using machine learning, International Organization for Standardization. https://www.iso.org/standard/74438.html. Accessed 2 June 2025

Mitchell J, RH (1988) P of the 15th A. SIGPLAN-SIGACT, undefined 1988, The essence of ML, Dl.Acm.Org. Part F130193:28–46. https://doi.org/10.1145/73560.73563

Jabareen Y, YJI (2009) Journal of qualitative methods, undefined 2009, building a conceptual framework: philosophy, definitions, and procedure. Int J Qual Methods 8:49–62

Jiao L, Wang Y, Liu, Li L, Liu F, Ma W, Guo Y, Chen P, Yang S, Hou B, (2024) Causal inference meets deep learning: A comprehensive survey, Spj. Science. OrgL Jiao, Y Wang, X Liu, L Li, F Liu, W Ma, Y Guo, P Chen, S Yang, B HouResearch, 2024 spj.Science. Org. 7 pp. 2025. https://doi.org/10.34133/RESEARCH.0467.

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1:389–399. https://doi.org/10.1038/s42256-019-0088-2

Kasirzadeh A (2021) A. Smart, The use and misuse of counterfactuals in ethical machine learning, In: FAccT 2021 - Pro. 2021 ACM Conf Fairness, Accountability, Transpar. https://doi.org/10.1145/3442188.3445886.

Kim J, Im I (2023) Anthropomorphic response: understanding interactions between humans and artificial intelligence agents. Comput Human Behav. https://doi.org/10.1016/j.chb.2022.107512

Kitcher P, Salmon WC (1962) Scientific explanation, (7563BC) pp. 528.

Klare BF, Burge MJ, Klontz JC, Vorder Bruegge RW, Jain AK (2012) Face recognition performance: role of demographic information. IEEE Trans Inf Forensics Secur. https://doi.org/10.1109/TIFS.2012.2214212

Kotsoglou KN (2013) How to become an epistemic engineer: what shifts when we change the standard of proof? Law Probab Risk. https://doi.org/10.1093/lpr/mgt002

Kroes P, Meijers A (2001) The empirical turn in the philosophy of technology. https://philpapers.org/rec/KROTET. Accessed 30 August 2025

Kuhn T (2020) The structure of scientific revolutions. Knowl Postmodernism Hist Perspect. https://doi.org/10.2307/2090620

Lake BM, Baroni M (2023) Human-like systematic generalization through a meta-learning neural network. Nature. https://doi.org/10.1038/s41586-023-06668-3

Langford GO (2016) Engineering systems integration: Theory, metrics, and methods.

Lazar S, Stone J (2024) On the site of predictive justice. Nous 58(NULL):730–754. https://doi.org/10.1111/NOUS.12477

Leavitt K, Schabram K, Hariharan P, Barnes CM (2024) The machine hums! Addressing ontological and normative concerns regarding machine learning applications in organizational scholarship. Acad Manage Rev 49(2):448–451. https://doi.org/10.5465/AMR.2021.0166

Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2018) Fair, transparent, and accountable algorithmic decision-making processes: the premise, the proposed solutions, and the open challenges. Philos Technol. https://doi.org/10.1007/s13347-017-0279-x

Liability of Employed Engineers | National Society of Professional Engineers, National Society of Professional Engineers (2023). https://www.nspe.org/career-growth/professional-resources/liability-risk/liability-employed-engineers. Accessed 4 June 2025

Lipton ZC (2018a) The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue. https://doi.org/10.1145/3236386.3241340

Lipton ZC (2018b) The mythos of model interpretability. Commun ACM. https://doi.org/10.1145/3233231

Liu J, Wang T, PC-A in N, undefined (2023) On the need for a language describing distribution shifts: Illustrations on tabular datasets, Proceedings.Neurips.Cc. (nd). https://proceedings.neurips.cc/paper_files/paper/2023/hash/a134eaebd55b7406ff29cd75d5f1a622-Abstract-Datasets_and_Benchmarks.html. Accessed 4 June 2025

Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G (2019) Learning under Concept Drift: A Review. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2018.2876857

Lysak V (2024) A theoretical analysis of philosophical dimensions in engineering education. Journal of Contemporary Philosophical and Anthropological Studies 2:2977–5507. https://doi.org/10.59652/JCPAS.V2I3.307

Mackenzie A (2015) The production of prediction: what does machine learning want? Eur J Cult Stud. https://doi.org/10.1177/1367549415577384

Marcus G, Davis E (2019) Rebooting AI: Building artificial intelligence we can trust, https://books.google.com/books?hl=en&lr=&id=OmeEDwAAQBAJ&oi=fnd&pg=PA3&dq=Marcus+%26+Davis,+2019&ots=LyYTaOYSfS&sig=wKkhNoaM5IOgZuEpywGhtmJbkq0. Accessed 3 June 2025)

Mari L, Wilson M, Maul A (2023) Philosophical Perspectives on Measurement, 81–121. https://doi.org/10.1007/978-3-031-22448-5_4

Matta C (2022) Philosophical paradigms in qualitative research methods education: what is their pedagogical role? Scand J Educ Res. https://doi.org/10.1080/00313831.2021.1958372

McCance KR, Blanchard M (2024) Measuring the interdisciplinarity and collaboration perceptions of U.S. scientists, engineers, and educators. AERA Open. https://doi.org/10.1177/23328584231218952

McDermott D (1976) Artificial intelligence meets natural stupidity. ACM SIGART Bulletin. https://doi.org/10.1145/1045339.1045340

Michelfelder DP, McCarthy N, Goldberg DE (2016) Philosophy and engineering: reflections on practice, principles and process. IEEE Trans Educ. https://doi.org/10.1109/te.2016.2518840

Michelfelder DP, Doorn N (2020) The Routledge Handbook of the Philosophy of Engineering. https://doi.org/10.4324/9781315276502.

Mitcham C (1998) The Importance of Philosophy to Engineering, Teorema.

Mody CCM (2021) Steps toward a Philosophy of Engineering: Historico-Philosophical and Critical Essays by Carl Mitcham. Technol Cult. https://doi.org/10.1353/tech.2021.0090

Moraffah R, Karami M, Guo R, Raglin A, Liu H (2020) Causal interpretability for machine learning - problems, methods and evaluation. ACM SIGKDD Explor Newsl. https://doi.org/10.1145/3400051.3400058

Morison E, Florman SC (1989) The civilized engineer. Technol Cult. https://doi.org/10.2307/3105443

Nahar N, Zhou S, Lewis G, Kastner C (2022) Collaboration challenges in building ML-enabled systems: communication, documentation, engineering, and process. Proc - Int Conf Softw Eng. https://doi.org/10.1145/3510003.3510209

Naser M (2024) Causality and causal inference for engineers: beyond correlation, regression, prediction and artificial intelligence. Wires Data Min Knowl Discov. https://doi.org/10.1002/WIDM.1533

Naser MZ (2025) A guide to machine learning epistemic ignorance, hidden paradoxes, and other tensions. Wires Data Min Knowl Discov. https://doi.org/10.1002/WIDM.70038

Newman J (2024) Promoting interdisciplinary research collaboration: a systematic review, a critical literature review, and a pathway forward. Soc Epistemol. https://doi.org/10.1080/02691728.2023.2172694

Ng AY, Jordan MI (2002) On discriminative vs. Generative classifiers: A comparison of logistic regression and naive bayes. In: Adv Neural Inf Process Syst

NSPE Code of Ethics for engineers 1993 JOM https://doi.org/10.1007/BF03223280

US NTSB, Collision Between a Car Operating With Automated... - Google Scholar, 2017. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C2&q=Collision+Between+a+Car+Operating+With+Automated+Vehicle+Control+Systems+and+a+Tractor-Semitrailer+Truck+Near+Williston%2C+Florida+May+7%2C+2016&btnG= (accessed August 30, 2025).

O'Hara I (2022) Automated Epistemology: Bots, Computational Propaganda & Information Literacy Instruction. J Academic Librarianship. https://doi.org/10.1016/j.acalib.2022.102540

C. O'Neil (2016). Weapons of math destruction: How big data increases inequality and threatens democracy, Crown

Oneto L, Ridella S, Anguita D (2024) Towards algorithms and models that we can trust: a theoretical perspective. Neurocomputing 592:127798. https://doi.org/10.1016/J.NEUCOM.2024.127798

Oreskes N, Shrader-Frechette K, Belitz K (1994) Verification, validation, and confirmation of numerical models in the Earth sciences. Science. https://doi.org/10.1126/science.263.5147.641

Wijesinghe P, Parker H, Stein E (2025) A Case Study: Students' Perception of the Use of Generative AI in Learning and the Civil Engineering Profession. American Society for Engineering Education. In: 2025 ASEE Annu Conf Expo https://nemo.asee.org/public/conferences/365/papers/47568/view? Accessed 3 October 2025

Parasuraman R, Riley V (1997) Humans and automation: use, misuse, disuse, abuse. Hum Factors. https://doi.org/10.1518/001872097778543886

van de Poel I, Royakkers L, Verbeek P-P, Brumsen M (2011) Ethics, technology and engineering: an introduction, Wiley-Blackwell, a John Wiley & Sons, Ltd., Publication

Popper K (2005) The Logic of Scientific Discovery. https://doi.org/10.4324/9780203994627

Ratcliff R (2013) Applying epistemology to system engineering: an illustration. Procedia Comput Sci. https://doi.org/10.1016/j.procs.2013.01.041

Rismani S, Shelby R, Smart A, Jatho E, Kroll J, Moon Aj, Rostamzadeh N (2023) From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. In: Conf Hum Factors Comput Syst - Proc. https://doi.org/10.1145/3544548.3581407.

Rossoni AL, de Vasconcellos EPG, de Castilho Rossoni RL (2024) Barriers and facilitators of university-industry collaboration for research, development and innovation: a systematic review. Manag Rev Q. https://doi.org/10.1007/s11301-023-00349-1

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. https://doi.org/10.1038/s42256-019-0048-x

Russell S, Norvig P (2010). Artificial Intelligence A Modern Approach Third Edition. https://doi.org/10.1017/S0269888900007724

Rybnicek R, Königsgruber R (2019) What makes industry–university collaboration succeed? A systematic review of the literature. J Bus Econ. https://doi.org/10.1007/s11573-018-0916-6

Sadek M, Calvo RA, Mougenot C (2023) Designing value-sensitive AI: a critical review and recommendations for socio-technical design processes. AI Ethics. https://doi.org/10.1007/s43681-023-00373-7

Salles A, Evers K, Farisco M (2020) Anthropomorphism in AI. AJOB Neurosci 11:88–95. https://doi.org/10.1080/21507740.2020.1740350

Schön DA 2017 The Reflective Practitioner: How Professionals Think in Action . https://doi.org/10.4324/9781315237473

Schuette R (1999) Architectures for Evaluating the Quality of Information Models — A Meta and an Object Level Comparison. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 1728:490–505. https://doi.org/10.1007/3-540-47866-3_33

Schweizer KW (2022) Artificial unintelligence: how computers misunderstand the world. Eur Legac. https://doi.org/10.1080/10848770.2022.2110366

Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems, In: FAT* 2019 - Proc 2019 conf fairness accountability, Transpar https://doi.org/10.1145/3287560.3287598.

Shafer G, Vovk V (2008) A tutorial on conformal prediction, Journal of Machine Learning Research.

Skorton D, Bear A (2018) N.A. of S. national academies, N.A. of E. national academies, N.A. of M. national academies, the integration of the humanities and arts with sciences, engineering, and medicine in higher education: branches from the same tree. Consensus Study Report

Slota SC, Fleischmann KR, Greenberg S, Verma N, Cummings B, Li L, Shenefiel C (2020) Good systems, bad data?: Interpretations of AI hype and failures. Proc Assoc Inf Sci Technol. https://doi.org/10.1002/pra2.275

Tereshchenko E, Salmela E, Melkko E, Phang SK, Happonen A (2024) Emerging best strategies and capabilities for university–industry

cooperation: opportunities for MSMEs and universities to improve collaboration. A Literature Review, SpringerE Tereshchenko, E Salmela, E Melkko, SK Phang, A HapponenJournal of Innovation and Entrepreneurship, Springer. 13:28. https://doi.org/10.1186/S13731-024-00386-4

CA-Trans/Form/Ação, undefined (2024) Behind artificial intelligence: an analysis of the epistemological bases of machine learning, SciELO BrasilC ArãoTrans/Form/Ação, 2024•SciELO Brasil. (n.d.). https://www.scielo.br/j/trans/a/wKP3thTz35fmhG9pnmXNdMj/abstract/?format=html&lang=en (accessed June 2, 2025).

Tulabandhula T, Rudin C (2014) On combining machine learning with decision making. Mach Learn. https://doi.org/10.1007/s10994-014-5459-7

Umbrello S, van de Poel I (2021) Mapping value sensitive design onto AI for social good principles. AI Ethics. https://doi.org/10.1007/s43681-021-00038-3

Vallor S (2020) The Oxford Handbook of Philosophy of Technology. https://doi.org/10.1093/oxfordhb/9780190851187.001.0001

van Giffen B, Herhausen D, Fahse T (2022) Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. J Bus Res. https://doi.org/10.1016/j.jbusres.2022.01.076

Varshney KR, Alemzadeh H (2017) On the safety of machine learning: cyber-physical systems, decision sciences, and data products. Big Data. https://doi.org/10.1089/big.2016.0051

Verma S, Boonsanong V, Hoang M, Hines K, Dickerson J, Shah C (2024) Counterfactual explanations and algorithmic recourses for machine learning: A review. ACM Comput Surv. https://doi.org/10.1145/3677119/ASSET/2907A6DC-DD42-4AF6-BE48-CA8C296851FC/ASSETS/GRAPHIC/CSUR-2023-0595-T02.JPG

Vermaas P, Didier C, Hanks C, Newberry B (2010) Philosophy of Engineering and Technology. https://www.academia.edu/download/109084299/bfm_978-3-319-16169-3_2F1.pdf. Accessed 3 June 2025

Waytz A, Heafner J, Epley N (2014) The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. J Exp Soc Psychol. https://doi.org/10.1016/j.jesp.2014.01.005

Wheeler G (2017) Machine epistemology and big data, The Routledge Companion to Philosophy of Social Science.

Wieringa M (2020) What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability, In: FAT* 2020 - Proc. 2020 conf. fairness, accountability, Transpar. https://doi.org/10.1145/3351095.3372833.

Wilson P (1995) Thinking through technology: the path between engineering and philosophy (book review). Coll Res Libr. https://doi.org/10.5860/crl_56_02_184

Winfield AFT, Jirotka M (2018) Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philos Trans R Soc Lond A Math Phys Eng Sci. https://doi.org/10.1098/rsta.2018.0085

Winner L (2020) The whale and the reactor: A search for limits in an age of high technology. https://books.google.com/books?hl=en&lr=&id=wKnQDwAAQBAJ&oi=fnd&pg=PR5&dq=The+whale+and+the+reactor:+A+search+for+limits+in+an+age+of+high+technology&ots=EEZ19br4e_&sig=9KSloy4SI1734m61nzVNYa5XR6M. Accessed 30 August 2025

Ye W, Zheng G, Cao X, Ma Y, Hu X, Zhang A (2024) Spurious correlations in machine learning: A survey. https://arxiv.org/pdf/2402.12715. Accessed 7 May 2025

Zafar MR, Khan N (2021) Deterministic local interpretable model-agnostic explanations for stable explainability. Mach Learn Knowl Extr. https://doi.org/10.3390/make3030027

Zemmouchi-Ghomari L (2023) Ontology and machine learning: A two-way street to improved knowledge representation and algorithm accuracy, 181–189. https://doi.org/10.1007/978-981-99-4626-6_15.

Zhang C, Kim J, Jeon J (2023) What does human-machine intelligence mean for civil... - Google Scholar, Civil Engineering Source Magazine. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C2&q=What+does+human-machine+intelligence+mean+for+civil+engineers%3F+%7C+ASCE&btnG=. Accessed 31 August 2025

Zhao Y, Wang Y, Derr T (2023) Fairness and explainability: bridging the gap towards fair model explanations. Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v37i9.26344

Žukauskas P, Vveinhardt J, Andriukaitienė R (2018) Philosophy and Paradigm of Scientific Research. Manag Cult Corp Soc Responsib. https://doi.org/10.5772/intechopen.70628

Zwart S (2022) Engineering epistemology: between theory and practice. Eng Stud. https://doi.org/10.1080/19378629.2022.2124025

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.